

Responsible AI

DoD's Ethical Principles for AI

Reliable

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Document Markings

Copyright 2021 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM21-1128

Reliable

Reliable

AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

Well-defined uses

Users and tasks are clear and situation is AI appropriate.

Proper data is available.

AI systems know ONLY what taught.

AI systems control ONLY what given control of.

AI systems can continue to learn, within that narrow scope.

What is the intention?

What problem are you solving?

For whom?

What will help them?

What kind of improvements?

What might a machine do better or faster?

What is not going to be improved (out of scope)?

Aspirations don't always work out

AI systems can learn bias embedded in data.

AI can continue to perpetuate discrimination against entire groups of individuals – extending historical norms.

Recent examples include:

- AI for financial and mortgage systems
- AI to determining likelihood of future behavior
- AI for facial recognition

Robust and secure for AI

Make it your business to protect people (and their data).

- Speculate to pre-emptively identify potential negative outcomes.
- Continuously conduct test and evaluation activities.

Make contingency plans.

- Identify warning signs for abuse/misuse.
- Plan to mitigate failures of the system quickly.
- Plan to monitor system continuously.

Security for AI

How do you secure the system from the inside and outside?

- Show examples of things going wrong
- Provide examples for when bias has changed, has increased
- Where do these examples live?

“If it’s not usable, it’s not secure.”
– Jared Spool, IAS17

Quote by Jared Spool. Unintuitive and Insecure: Fixing the Failures of Authentication, IA Summit 2017

Use cases

Consider for the following use cases

- What knowledge needed?
- What is it perceiving?
- What are ethical considerations?
- What are the limitations?

Strategic games

1997 Chess, IBM

2016 Go, Google

Knowledge?

Perception?

Ethics?

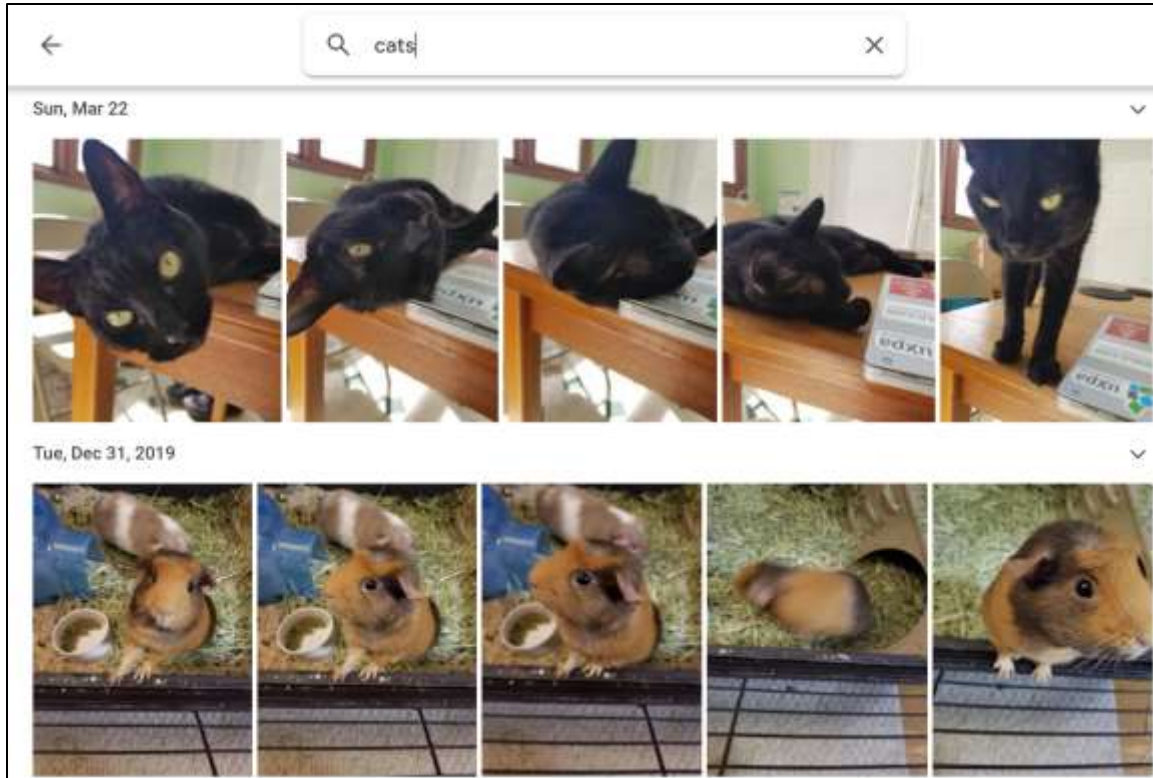
Limitations?



Floor goban, 2007, By Goban1

<https://commons.wikimedia.org/wiki/File:FloorGoban.JPG>

Image recognition – Google Photos



Assess natural disaster building damage



NASA - Fighting Fires Together: xView 2 Prize Challenge Helps Automate Damage Assessments

<https://appliedsciences.nasa.gov/our-impact/news/fighting-fires-together-xview-2-prize-challenge-helps-automate-damage-assessments>

Sound recognition: labeling of birdsongs



Photo by Gallo71 (Own work) [Public domain], via Wikimedia Commons
<https://commons.wikimedia.org/wiki/File%3ARbruni.JPG>

Listening and understanding human speech

Mapping Q & A + AI

Expected language

Appropriate automated responses

When to escalate?

- Searches on self harm?
- What else?



Hi, I'm Woebot!



Images: <https://www.pexels.com/photo/close-up-of-mobile-phone-248512/>
<https://www.amazon.com/Amazon-Echo-Bluetooth-Speaker-with-WiFi-Alexa/dp/B00X4WHP5E>

Decision making: autonomous vehicles



Image: <https://www.uber.com/info/atg/>

Reliability across lifecycle

Speculation saves resources.

Fixing issues later not always possible.

Test and evaluation efforts must be continuous.

Reliable – Review

Explicit, well-defined uses that are AI appropriate.

Safety, security, and effectiveness tested across entire life-cycle.

**Empower diverse teams in
inclusive environments.**

**Encourage deep conversations, speculation,
and imaginative thinking.**

Alex Van Deusen
Design Researcher
arvandeusen@sei.cmu.edu

Carol Smith
Sr. Research Scientist
cjsmith@sei.cmu.edu

Rachel Dzombak
Digital Transformation Lead
rdzombak@sei.cmu.edu